CHAPTER 12

12.1. Take the conditional expectation of equation (12.4) with respect to $\mathbf{x}$, and use $\mathrm{E}(u|\mathbf{x}) = 0$:

$$\begin{aligned}\mathrm{E}\{[y - m(\mathbf{x},\boldsymbol{\theta})]^2|\mathbf{x}\} &= \mathrm{E}(u^2|\mathbf{x}) + 2[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]\mathrm{E}(u|\mathbf{x}) \\ &\quad + \mathrm{E}\{[m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2|\mathbf{x}\} \\ &= \mathrm{E}(u^2|\mathbf{x}) + 0 + [m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2 \\ &= \mathrm{E}(u^2|\mathbf{x}) + [m(\mathbf{x},\boldsymbol{\theta}_o) - m(\mathbf{x},\boldsymbol{\theta})]^2.\end{aligned}$$

Now, the first term does not depend on $\boldsymbol{\theta}$, and the second term is clearly minimized at $\boldsymbol{\theta} = \boldsymbol{\theta}_o$ (although not uniquely, in general).

12.3. a. The approximate elasticity is $\partial\log[\hat{\mathrm{E}}(y|\mathbf{z})]/\partial\log(z_1) = \partial[\hat{\theta}_1 + \hat{\theta}_2\log(z_1) + \hat{\theta}_3 z_2]/\partial\log(z_1) = \hat{\theta}_2$.

b. This is approximated by $100\cdot\partial\log[\hat{\mathrm{E}}(y|\mathbf{z})]/\partial z_2 = 100\cdot\hat{\theta}_3$.

c. Since $\partial\hat{\mathrm{E}}(y|\mathbf{z})/\partial z_2 = \exp[\hat{\theta}_1 + \hat{\theta}_2\log(z_1) + \hat{\theta}_3 z_2 + \hat{\theta}_4 z_2^2]\cdot(\hat{\theta}_3 + 2\hat{\theta}_4 z_2)$, the turning point is $z_2^* = \hat{\theta}_3/(-2\hat{\theta}_4)$.

d. Since $\nabla_\theta m(\mathbf{x},\boldsymbol{\theta}) = \exp(\mathbf{x}_1\boldsymbol{\theta}_1 + \mathbf{x}_2\boldsymbol{\theta}_2)\mathbf{x}$, the gradient of the mean function evaluated under the null is $\nabla_\theta\tilde{m}_i = \exp(\mathbf{x}_{i1}\tilde{\boldsymbol{\theta}}_1)\mathbf{x}_i \equiv \tilde{m}_i\mathbf{x}_i$, where $\tilde{\boldsymbol{\theta}}_1$ is the restricted NLS estimator. From (12.72), we can compute the usual LM statistic as NR_u^2 from the regression $\tilde{u}_i$ on $\tilde{m}_i\mathbf{x}_{i1}$, $\tilde{m}_i\mathbf{x}_{i2}$, $i = 1,\ldots,N$, where $\tilde{u}_i = y_i - \tilde{m}_i$. For the robust test, we first regress $\tilde{m}_i\mathbf{x}_{i2}$ on $\tilde{m}_i\mathbf{x}_{i1}$ and obtain the $1 \times K_2$ residuals, $\tilde{\mathbf{r}}_i$. Then we compute the statistic as in regression (12.75).

12.5. We need the gradient of $m(\mathbf{x}_i,\boldsymbol{\theta})$ evaluated under the null hypothesis. By the chain rule, $\nabla_\beta m(\mathbf{x},\boldsymbol{\theta}) = g[\mathbf{x}\boldsymbol{\beta} + \delta_1(\mathbf{x}\boldsymbol{\beta})^2 + \delta_2(\mathbf{x}\boldsymbol{\beta})^3]\cdot[\mathbf{x} + 2\delta_1(\mathbf{x}\boldsymbol{\beta})^2 +$

$3\delta_2(\mathbf{x}\boldsymbol{\beta})^2]$, $\nabla_\delta m(\mathbf{x},\boldsymbol{\theta}) = g[\mathbf{x}\boldsymbol{\beta} + \delta_1(\mathbf{x}\boldsymbol{\beta})^2 + \delta_2(\mathbf{x}\boldsymbol{\beta})^3]\cdot[(\mathbf{x}\boldsymbol{\beta})^2,(\mathbf{x}\boldsymbol{\beta})^3]$. Let $\tilde{\boldsymbol{\beta}}$ denote the NLS estimator with $\delta_1 = \delta_2 = 0$ imposed. Then $\nabla_\beta m(\mathbf{x}_i,\tilde{\boldsymbol{\theta}}) = g(\mathbf{x}_i\tilde{\boldsymbol{\beta}})\mathbf{x}_i$ and $\nabla_\delta m(\mathbf{x}_i,\tilde{\boldsymbol{\theta}}) = g(\mathbf{x}_i\tilde{\boldsymbol{\beta}})[(\mathbf{x}_i\tilde{\boldsymbol{\beta}})^2,(\mathbf{x}_i\tilde{\boldsymbol{\beta}})^3]$. Therefore, the usual LM statistic can be obtained as NR_u^2 from the regression $\tilde{u}_i$ on $\tilde{g}_i\mathbf{x}_i$, $\tilde{g}_i\cdot(\mathbf{x}_i\tilde{\boldsymbol{\beta}})^2$, $\tilde{g}_i\cdot(\mathbf{x}_i\tilde{\boldsymbol{\beta}})^3$, where $\tilde{g}_i \equiv g(\mathbf{x}_i\tilde{\boldsymbol{\beta}})$. If $G(\cdot)$ is the identity function, $g(\cdot) \equiv 1$, and we get RESET.

12.7. a. For each i and g, define $u_{ig} \equiv y_{ig} - m(\mathbf{x}_{ig},\boldsymbol{\theta}_o)$, so that $E(u_{ig}|\mathbf{x}_i) = 0$, $g = 1,\ldots,G$. Further, let $\mathbf{u}_i$ be the $G \times 1$ vector containing the u_{ig}. Then $E(\mathbf{u}_i\mathbf{u}_i'|\mathbf{x}_i) = E(\mathbf{u}_i\mathbf{u}_i') = \boldsymbol{\Omega}_o$. Let $\hat{\hat{\mathbf{u}}}_i$ be the vector of nonlinear least squares residuals. That is, do NLS for each g, and collect the residuals. Then, by standard arguments, a consistent estimator of $\boldsymbol{\Omega}_o$ is

$$\hat{\boldsymbol{\Omega}} \equiv N^{-1}\sum_{i=1}^{N}\hat{\hat{\mathbf{u}}}_i\hat{\hat{\mathbf{u}}}_i'$$

because each NLS estimator, $\hat{\hat{\boldsymbol{\theta}}}_g$ is consistent for $\boldsymbol{\theta}_{og}$ as $N \to \infty$.

b. This part involves several steps, and I will sketch how each one goes. First, let $\boldsymbol{\gamma}$ be the vector of distinct elements of $\boldsymbol{\Omega}$ -- the nuisance parameters in the context of two-step M-estimation. Then, the score for observation i is

$$\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta};\boldsymbol{\gamma}) = -\nabla_\theta \mathbf{m}(\mathbf{x}_i,\boldsymbol{\theta})'\boldsymbol{\Omega}^{-1}\mathbf{u}_i(\boldsymbol{\theta})$$

where, hopefully, the notation is clear. With this definition, we can verify condition (12.37), even though the actual derivatives are complicated. Each element of $\mathbf{s}(\mathbf{w}_i,\boldsymbol{\theta};\boldsymbol{\gamma})$ is a linear combination of $\mathbf{u}_i(\boldsymbol{\theta})$. So $\nabla_\gamma s_j(\mathbf{w}_i,\boldsymbol{\theta}_o;\boldsymbol{\gamma})$ is a linear combination of $\mathbf{u}_i(\boldsymbol{\theta}_o) \equiv \mathbf{u}_i$, where the linear combination is a function of $(\mathbf{x}_i,\boldsymbol{\theta}_o,\boldsymbol{\gamma})$. Since $E(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$, $E[\nabla_\gamma s_j(\mathbf{w}_i,\boldsymbol{\theta}_o;\boldsymbol{\gamma})|\mathbf{x}_i] = \mathbf{0}$, and so its unconditional expectation is zero, too. This shows that we do not have to adjust for the first-stage estimation of $\boldsymbol{\Omega}_o$. Alternatively, one can verify the hint directly, which has the same consequence.

Next, we derive $\mathbf{B}_o \equiv E[\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}_o)\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}_o)']$:

$$
\begin{aligned}
E[\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}_o)\mathbf{s}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}_o)'] &= E[\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\mathbf{u}_i\mathbf{u}_i'\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)] \\
&= E\{E[\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\mathbf{u}_i\mathbf{u}_i'\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)|\mathbf{x}_i]\} \\
&= E[\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}E(\mathbf{u}_i\mathbf{u}_i'|\mathbf{x}_i)\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)] \\
&= E[\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\boldsymbol{\Omega}_o\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)] = E[\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)].
\end{aligned}
$$

Next, we have to derive $\mathbf{A}_o \equiv E[\mathbf{H}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}_o)]$, and show that $\mathbf{B}_o = \mathbf{A}_o$. The Hessian itself is complicated, but its expected value is not. The Jacobian of $\mathbf{s}_i(\boldsymbol{\theta};\boldsymbol{\gamma})$ with respect to $\boldsymbol{\theta}$ can be written

$$
\mathbf{H}_i(\boldsymbol{\theta};\boldsymbol{\gamma}) = \nabla_\theta\mathbf{m}(\mathbf{x}_i,\boldsymbol{\theta})'\boldsymbol{\Omega}^{-1}\nabla_\theta\mathbf{m}(\mathbf{x}_i,\boldsymbol{\theta}) + [\mathbf{I}_P \otimes \mathbf{u}_i(\boldsymbol{\theta})']\mathbf{F}(\mathbf{x}_i,\boldsymbol{\theta};\boldsymbol{\gamma}),
$$

where $\mathbf{F}(\mathbf{x}_i,\boldsymbol{\theta};\boldsymbol{\gamma})$ is a $GP \times P$ matrix, where P is the total number of parameters, that involves Jacobians of the rows of $\boldsymbol{\Omega}^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The key is that $\mathbf{F}(\mathbf{x}_i,\boldsymbol{\theta};\boldsymbol{\gamma})$ depends on $\mathbf{x}_i$, not on $\mathbf{y}_i$. So,

$$
\begin{aligned}
E[\mathbf{H}_i(\boldsymbol{\theta}_o;\boldsymbol{\gamma}_o)|\mathbf{x}_i] &= \nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o) + [\mathbf{I}_P \otimes E(\mathbf{u}_i|\mathbf{x}_i)']\mathbf{F}(\mathbf{x}_i,\boldsymbol{\theta}_o;\boldsymbol{\gamma}_o) \\
&= \nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o).
\end{aligned}
$$

Now iterated expectations gives $\mathbf{A}_o = E[\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)]$. So, we have verified (12.37) and that $\mathbf{A}_o = \mathbf{B}_o$. Therefore, from Theorem 12.3, Avar $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o) = \mathbf{A}_o^{-1} = \{E[\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)]\}^{-1}$.

c. As usual, we replace expectations with sample averages and unknown parameters, and divide the result by N to get $\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}})$:

$$
\begin{aligned}
\widehat{\text{Avar}}(\hat{\boldsymbol{\theta}}) &= \left(N^{-1}\sum_{i=1}^{N}\nabla_\theta\mathbf{m}_i(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Omega}}^{-1}\nabla_\theta\mathbf{m}_i(\hat{\boldsymbol{\theta}})\right)^{-1}/N \\
&= \left(\sum_{i=1}^{N}\nabla_\theta\mathbf{m}_i(\hat{\boldsymbol{\theta}})'\hat{\boldsymbol{\Omega}}^{-1}\nabla_\theta\mathbf{m}_i(\hat{\boldsymbol{\theta}})\right)^{-1}.
\end{aligned}
$$

The estimate $\hat{\boldsymbol{\Omega}}$ can be based on the multivariate NLS residuals or can be updated after the nonlinear SUR estimates have been obtained.

d. First, note that $\nabla_\theta\mathbf{m}_i(\boldsymbol{\theta}_o)$ is a block-diagonal matrix with blocks $\nabla_{\theta_g}m_{ig}(\boldsymbol{\theta}_{og})$, a $1 \times P_g$ matrix. (I implicityly assumed that there are no cross-equation restrictions imposed in the nonlinear SUR estimation.) If $\boldsymbol{\Omega}_o$

is block-diagonal, so is its inverse. Standard matrix multiplication shows that

$$\nabla_\theta \mathbf{m}_i(\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o^{-1}\nabla_\theta \mathbf{m}_i(\boldsymbol{\theta}_o) = \begin{pmatrix} \sigma_{o1}^{-2}\nabla_{\theta_1} m_{i1}^o{}'\nabla_{\theta_1} m_{i1}^o & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & & & \\ \vdots & & & \\ \mathbf{0} & \cdots & \mathbf{0} & \sigma_{oG}^{-2}\nabla_{\theta_G} m_{iG}^o{}'\nabla_{\theta_G} m_{iG}^o \end{pmatrix}.$$

Taking expectations and inverting the result shows that Avar $\sqrt{N}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_{og}) = \sigma_{og}^2[\mathrm{E}(\nabla_{\theta_g} m_{ig}^o{}'\nabla_{\theta_g} m_{ig}^o)]^{-1}$, $g = 1,\ldots,G$. (Note also that the nonlinear SUR estimators are asymptotically uncorrelated across equations.) These asymptotic variances are easily seen to be the same as those for nonlinear least squares on each equation; see p. 360.

e. I cannot see a nonlinear analogue of Theorem 7.7. The first hint given in Problem 7.5 does not extend readily to nonlinear models, even when the same regressors appear in each equation. The key is that $\mathbf{X}_i$ is replaced with $\nabla_\theta \mathbf{m}(\mathbf{x}_i,\boldsymbol{\theta}_o)$. While this $G \times P$ matrix has a block-diagonal form, as described in part d, the blocks are not the same even when the same regressors appear in each equation. In the linear case, $\nabla_{\theta_g} m_g(\mathbf{x}_i,\boldsymbol{\theta}_{og}) = \mathbf{x}_i$ for all g. But, unless $\boldsymbol{\theta}_{og}$ is the same in all equations -- a very restrictive assumption -- $\nabla_{\theta_g} m_g(\mathbf{x}_i,\boldsymbol{\theta}_{og})$ varies across g. For example, if $m_g(\mathbf{x}_i,\boldsymbol{\theta}_{og}) = \exp(\mathbf{x}_i\boldsymbol{\theta}_{og})$ then $\nabla_{\theta_g} m_g(\mathbf{x}_i,\boldsymbol{\theta}_{og}) = \exp(\mathbf{x}_i\boldsymbol{\theta}_{og})\mathbf{x}_i$, and the gradients differ across g.

12.9. a. We cannot say anything in general about $\mathrm{Med}(y|\mathbf{x})$, since $\mathrm{Med}(y|\mathbf{x}) = m(\mathbf{x},\boldsymbol{\beta}_o) + \mathrm{Med}(u|\mathbf{x})$, and $\mathrm{Med}(u|\mathbf{x})$ could be a general function of $\mathbf{x}$.

b. If u and $\mathbf{x}$ are independent, then $\mathrm{E}(u|\mathbf{x})$ and $\mathrm{Med}(u|\mathbf{x})$ are both constants, say α and δ. Then $\mathrm{E}(y|\mathbf{x}) - \mathrm{Med}(y|\mathbf{x}) = \alpha - \delta$, which does not depend on $\mathbf{x}$.

c. When u and $\mathbf{x}$ are independent, the partial effects of x_j on the conditional mean and conditional median are the same, and there is no ambiguity about what is "the effect of x_j on y," at least when only the mean and median are in the running. Then, we could interpret large differences between LAD and NLS as perhaps indicating an outlier problem. But it could just be that u and $\mathbf{x}$ are not independent.

12.11. a. For consistency of the MNLS estimator, we need -- in addition to the regularity conditions, which I will ignore -- the identification condition. That is, $\boldsymbol{\beta}_o$ must uniquely minimize $\mathrm{E}[q(\mathbf{w}_i,\boldsymbol{\beta})] = \mathrm{E}\{[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\} = \mathrm{E}(\{\mathbf{u}_i + [\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\}'\{\mathbf{u}_i + [\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\}) = \mathrm{E}(\mathbf{u}_i'\mathbf{u}_i) + 2\mathrm{E}\{[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'\mathbf{u}_i\} + \mathrm{E}\{[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\} = \mathrm{E}(\mathbf{u}_i\mathbf{u}_i') + \mathrm{E}\{[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\}$ because $\mathrm{E}(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$. Therefore, the identification assumption is that

$$\mathrm{E}\{[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\} > 0, \ \boldsymbol{\beta} \neq \boldsymbol{\beta}_o.$$

In a linear model, where $\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}) = \mathbf{X}_i\boldsymbol{\beta}$ for $\mathbf{X}_i$ a $G \times K$ matrix, the condition is

$$(\boldsymbol{\beta}_o - \boldsymbol{\beta})'\mathrm{E}(\mathbf{X}_i'\mathbf{X}_i)(\boldsymbol{\beta}_o - \boldsymbol{\beta}) > 0, \ \boldsymbol{\beta} \neq \boldsymbol{\beta}_o,$$

and this holds provided $\mathrm{E}(\mathbf{X}_i'\mathbf{X}_i)$ is positive definite.

Provided $\mathbf{m}(\mathbf{x},\cdot)$ is twice continuously differentiable, there are no problems in applying Theorem 12.3. Generally, $\mathbf{B}_o = \mathrm{E}[\nabla_\theta\mathbf{m}_i(\boldsymbol{\beta}_o)'\mathbf{u}_i\mathbf{u}_i'\nabla_\theta\mathbf{m}_i(\boldsymbol{\beta}_o)]$ and $\mathbf{A}_o = \mathrm{E}[\nabla_\theta\mathbf{m}_i(\boldsymbol{\beta}_o)'\nabla_\theta\mathbf{m}_i(\boldsymbol{\beta}_o)]$. These can be consistently estimated in the obvious way after obtain the MNLS estimators.

b. We can apply the results on two-step M-estimation. The key is that, underl general regularity conditions,

$$N^{-1}\sum_{i=1}^{N}[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{W}_i(\hat{\boldsymbol{\delta}})]^{-1}[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]/2,$$

converges uniformly in probability to

$$\mathrm{E}\{[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{W}(\mathbf{x}_i,\boldsymbol{\delta}_o)]^{-1}[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\}/2,$$

which is just to say that the usual consistency proof can be used provided we verify identification. But we can use an argument very similar to the unweighted case to show

$$\mathrm{E}\{[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{W}(\mathbf{x}_i,\boldsymbol{\delta}_o)]^{-1}[\mathbf{y}_i - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\} = \mathrm{E}\{\mathbf{u}_i'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\mathbf{u}_i\}$$
$$+ \mathrm{E}\{[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]\},$$

where $\mathrm{E}(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{0}$ is used to show the cross-product term, $2\mathrm{E}\{[\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) - \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta})]'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\mathbf{u}_i\}$, is zero (by iterated expectations, as always). As before, the first term does not depend on $\boldsymbol{\beta}$ and the second term is minimized at $\boldsymbol{\beta}_o$; we would have to assume it is uniquely minimized.

To get the asymptotic variance, we proceed as in Problem 12.7. First, it can be shown that condition (12.37) holds. In particular, we can write $\nabla_{\delta}\mathbf{s}_i(\boldsymbol{\beta}_o;\boldsymbol{\delta}_o) = (\mathbf{I}_P \otimes \mathbf{u}_i)'\mathbf{G}(\mathbf{x}_i,\boldsymbol{\beta}_o;\boldsymbol{\delta}_o)$ for some function $\mathbf{G}(\mathbf{x}_i,\boldsymbol{\beta}_o;\boldsymbol{\delta}_o)$. It follows easily that $\mathrm{E}[\nabla_{\delta}\mathbf{s}_i(\boldsymbol{\beta}_o;\boldsymbol{\delta}_o)|\mathbf{x}_i] = \mathbf{0}$, which implies (12.37). This means that, under $\mathrm{E}(\mathbf{y}_i|\mathbf{x}_i) = \mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o)$, we can ignore preliminary estimation of $\boldsymbol{\delta}_o$ provided we have a $\sqrt{N}$-consistent estimator.

To obtain the asymptotic variance when the conditional variance matrix is correctly specified, that is, when $\mathrm{Var}(\mathbf{y}_i|\mathbf{x}_i) = \mathrm{Var}(\mathbf{u}_i|\mathbf{x}_i) = \mathbf{W}(\mathbf{x}_i,\boldsymbol{\delta}_o)$, we can use an argument very similar to the nonlinear SUR case in Problem 12.7:

$$\mathrm{E}[\mathbf{s}_i(\boldsymbol{\beta}_o;\boldsymbol{\delta}_o)\mathbf{s}_i(\boldsymbol{\beta}_o;\boldsymbol{\delta}_o)'] = \mathrm{E}[\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\mathbf{u}_i\mathbf{u}_i'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)]$$
$$= \mathrm{E}\{\mathrm{E}[\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\mathbf{u}_i\mathbf{u}_i'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)|\mathbf{x}_i]\}$$
$$= \mathrm{E}[\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\mathrm{E}(\mathbf{u}_i\mathbf{u}_i'|\mathbf{x}_i)[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)]$$
$$= \mathrm{E}\{\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}]\nabla_{\beta}\mathbf{m}_i(\boldsymbol{\beta}_o)\}.$$

Now, the Hessian (with respect to $\boldsymbol{\beta}$), evaluated at $(\boldsymbol{\beta}_o,\boldsymbol{\delta}_o)$, can be written as

$$\mathbf{H}_i(\boldsymbol{\beta}_o;\boldsymbol{\delta}_o) = \nabla_\beta\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o)'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\nabla_\beta\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o) + (\mathbf{I}_P \otimes \mathbf{u}_i)']\mathbf{F}(\mathbf{x}_i,\boldsymbol{\beta}_o;\boldsymbol{\delta}_o),$$

for some complicated function $\mathbf{F}(\mathbf{x}_i,\boldsymbol{\beta}_o;\boldsymbol{\delta}_o)$ that depends only on $\mathbf{x}_i$. Taking expectations gives

$$\mathbf{A}_o \equiv \mathrm{E}[\mathbf{H}_i(\boldsymbol{\beta}_o;\boldsymbol{\delta}_o)] = \mathrm{E}\{\nabla_\beta\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o)'[\mathbf{W}_i(\boldsymbol{\delta}_o)]^{-1}\nabla_\beta\mathbf{m}(\mathbf{x}_i,\boldsymbol{\beta}_o)\} = \mathbf{B}_o.$$

Therefore, from the usual results on M-estimation, Avar $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o) = \mathbf{A}_o^{-1}$, and a consistent estimator of $\mathbf{A}_o$ is

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N}\nabla_\beta\mathbf{m}(\mathbf{x}_i,\hat{\boldsymbol{\beta}})'[\mathbf{W}_i(\hat{\boldsymbol{\delta}})]^{-1}\nabla_\beta\mathbf{m}(\mathbf{x}_i,\hat{\boldsymbol{\beta}}).$$

c. The consistency argument in part b did not use the fact that $\mathbf{W}(\mathbf{x},\boldsymbol{\delta})$ is correctly specified for $\mathrm{Var}(\mathbf{y}|\mathbf{x})$. Exactly the same derivation goes through. But, of course, the asymptotic variance is affected because $\mathbf{A}_o \neq \mathbf{B}_o$, and the expression for $\mathbf{B}_o$ no longer holds. The estimator of $\mathbf{A}_o$ in part b still works, of course. To consistently estimate $\mathbf{B}_o$ we use

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\nabla_\beta\mathbf{m}(\mathbf{x}_i,\hat{\boldsymbol{\beta}})'[\mathbf{W}_i(\hat{\boldsymbol{\delta}})]^{-1}\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'[\mathbf{W}_i(\hat{\boldsymbol{\delta}})]^{-1}\nabla_\beta\mathbf{m}(\mathbf{x}_i,\hat{\boldsymbol{\beta}}).$$

Now, we estimate Avar $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_o)$ in the usual way: $\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$.

CHAPTER 13

13.1. No. We know that $\boldsymbol{\theta}_o$ solves

$$\max_{\boldsymbol{\theta}\in\Theta} \mathrm{E}[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})],$$

where the expectation is over the joint distribution of $(\mathbf{x}_i,\mathbf{y}_i)$. Therefore, because $\exp(\cdot)$ is an increasing function, $\boldsymbol{\theta}_o$ also maximizes $\exp\{\mathrm{E}[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})]\}$ over Θ. The problem is that the expectation and the exponential function cannot be interchanged: $\mathrm{E}[f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})] \neq \exp\{\mathrm{E}[\log f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})]\}$. In fact, Jensen's inequality tells us that $\mathrm{E}[f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})] > \exp\{\mathrm{E}[\log$

$f(\mathbf{y}_i|\mathbf{x}_i;\boldsymbol{\theta})]\}$.

13.3. Parts a and b essentially appear in Section 15.4.

13.5. a. Since $\mathbf{s}_i^g(\boldsymbol{\phi}_o) = [\mathbf{G}(\boldsymbol{\theta}_o)']^{-1}\mathbf{s}_i(\boldsymbol{\theta}_o)$,

$$\begin{aligned}\mathrm{E}[\mathbf{s}_i^g(\boldsymbol{\phi}_o)\mathbf{s}_i^g(\boldsymbol{\phi}_o)'|\mathbf{x}_i] &= \mathrm{E}\{[\mathbf{G}(\boldsymbol{\theta}_o)']^{-1}\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'[\mathbf{G}(\boldsymbol{\theta}_o)]^{-1}|\mathbf{x}_i\} \\ &= [\mathbf{G}(\boldsymbol{\theta}_o)']^{-1}\mathrm{E}[\mathbf{s}_i(\boldsymbol{\theta}_o)\mathbf{s}_i(\boldsymbol{\theta}_o)'|\mathbf{x}_i][\mathbf{G}(\boldsymbol{\theta}_o)]^{-1} \\ &= [\mathbf{G}(\boldsymbol{\theta}_o)']^{-1}\mathbf{A}_i(\boldsymbol{\theta}_o)[\mathbf{G}(\boldsymbol{\theta}_o)]^{-1}.\end{aligned}$$

b. In part b, we just replace $\boldsymbol{\theta}_o$ with $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\phi}_o$ with $\tilde{\boldsymbol{\phi}}$:

$$\tilde{\mathbf{A}}_i^g = [\mathbf{G}(\tilde{\boldsymbol{\theta}})']^{-1}\mathbf{A}_i(\tilde{\boldsymbol{\theta}})[\mathbf{G}(\tilde{\boldsymbol{\theta}})]^{-1} \equiv \tilde{\mathbf{G}}'^{-1}\tilde{\mathbf{A}}_i\tilde{\mathbf{G}}^{-1}.$$

c. The expected Hessian form of the statistic is given in the second part of equation (13.36), but where it is based initial on $\tilde{\mathbf{s}}_i^g$ and $\tilde{\mathbf{A}}_i^g$:

$$\begin{aligned}LM_g &= \left(\sum_{i=1}^N \tilde{\mathbf{s}}_i^g\right)'\left(\sum_{i=1}^N \tilde{\mathbf{A}}_i^g\right)^{-1}\left(\sum_{i=1}^N \tilde{\mathbf{s}}_i^g\right) \\ &= \left(\sum_{i=1}^N \tilde{\mathbf{G}}'^{-1}\tilde{\mathbf{s}}_i\right)'\left(\sum_{i=1}^N \tilde{\mathbf{G}}'^{-1}\tilde{\mathbf{A}}_i\tilde{\mathbf{G}}^{-1}\right)^{-1}\left(\sum_{i=1}^N \tilde{\mathbf{G}}'^{-1}\tilde{\mathbf{s}}_i\right) \\ &= \left(\sum_{i=1}^N \tilde{\mathbf{s}}_i\right)'\tilde{\mathbf{G}}^{-1}\tilde{\mathbf{G}}\left(\sum_{i=1}^N \tilde{\mathbf{A}}_i\right)^{-1}\tilde{\mathbf{G}}'\tilde{\mathbf{G}}'^{-1}\left(\sum_{i=1}^N \tilde{\mathbf{s}}_i\right) \\ &= \left(\sum_{i=1}^N \tilde{\mathbf{s}}_i\right)'\left(\sum_{i=1}^N \tilde{\mathbf{A}}_i\right)^{-1}\left(\sum_{i=1}^N \tilde{\mathbf{s}}_i\right) = LM.\end{aligned}$$

13.7. a. The joint density is simply $g(y_1|y_2,\mathbf{x};\boldsymbol{\theta}_o)\cdot h(y_2|\mathbf{x};\boldsymbol{\theta}_o)$. The log-likelihood for observation i is

$$\ell_i(\boldsymbol{\theta}) \equiv \log g(y_{i1}|y_{i2},\mathbf{x}_i;\boldsymbol{\theta}) + \log h(y_{i2}|\mathbf{x}_i;\boldsymbol{\theta}),$$

and we would use this in a standard MLE analysis (conditional on $\mathbf{x}_i$).

b. First, we know that, for all $(y_{i2},\mathbf{x}_i)$, $\boldsymbol{\theta}_o$ maximizes $\mathrm{E}[\ell_{i1}(\boldsymbol{\theta})|y_{i2},\mathbf{x}_i]$. Since r_{i2} is a function of $(y_{i2},\mathbf{x}_i)$,

$$\mathrm{E}[r_{i2}\ell_{i1}(\boldsymbol{\theta})|y_{i2},\mathbf{x}_i] = r_{i2}\mathrm{E}[\ell_{i1}(\boldsymbol{\theta})|y_{i2},\mathbf{x}_i];$$

since $r_{i2} \geq 1$, $\boldsymbol{\theta}_o$ maximizes $\mathrm{E}[r_{i2}\ell_{i1}(\boldsymbol{\theta})|y_{i2},\mathbf{x}_i]$ for all $(y_{i2},\mathbf{x}_i)$, and therefore $\boldsymbol{\theta}_o$ maximizes $\mathrm{E}[r_{i2}\ell_{i1}(\boldsymbol{\theta})]$. Similary, $\boldsymbol{\theta}_o$ maximizes $\mathrm{E}[\ell_{i2}(\boldsymbol{\theta})]$, and so it follows that $\boldsymbol{\theta}_{o}$ maximizes $r_{i2}\ell_{i1}(\boldsymbol{\theta}) + \ell_{i2}(\boldsymbol{\theta})$. For identification, we have to assume or verify uniqueness.

c. The score is

$$\mathbf{s}_{i}(\boldsymbol{\theta}) = r_{i2}\mathbf{s}_{i1}(\boldsymbol{\theta}) + \mathbf{s}_{i2}(\boldsymbol{\theta}),$$

where $\mathbf{s}_{i1}(\boldsymbol{\theta}) \equiv \nabla_{\theta}\ell_{i1}(\boldsymbol{\theta})'$ and $\mathbf{s}_{i2}(\boldsymbol{\theta}) \equiv \nabla_{\theta}\ell_{i2}(\boldsymbol{\theta})'$. Therefore,

$$\begin{aligned}\mathrm{E}[\mathbf{s}_{i}(\boldsymbol{\theta}_{o})\mathbf{s}_{i}(\boldsymbol{\theta}_{o})'] &= \mathrm{E}[r_{i2}\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})'] + \mathrm{E}[\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})'] \\ &\quad + \mathrm{E}[r_{i2}\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})'] + \mathrm{E}[r_{i2}\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})'].\end{aligned}$$

Now by the usual conditional MLE theory, $\mathrm{E}[\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})|y_{i2},\mathbf{x}_{i}] = \mathbf{0}$ and, since r_{i2} and $\mathbf{s}_{i2}(\boldsymbol{\theta})$ are functions of $(y_{i2},\mathbf{x}_{i})$, it follows that $\mathrm{E}[r_{i2}\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})'|y_{i2},\mathbf{x}_{i}] = \mathbf{0}$, and so its transpose also has zero conditional expectation. As usual, this implies zero unconditional expectation. We have shown

$$\mathrm{E}[\mathbf{s}_{i}(\boldsymbol{\theta}_{o})\mathbf{s}_{i}(\boldsymbol{\theta}_{o})'] = \mathrm{E}[r_{i2}\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})'] + \mathrm{E}[\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})'].$$

Now, by the unconditional information matrix equality for the density $h(y_{2}|\mathbf{x};\boldsymbol{\theta})$, $\mathrm{E}[\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})\mathbf{s}_{i2}(\boldsymbol{\theta}_{o})'] = -\mathrm{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_{o})]$, where $\mathbf{H}_{i2}(\boldsymbol{\theta}) = \nabla_{\theta}\mathbf{s}_{i2}(\boldsymbol{\theta})$. Further, byt the conditional IM equality for the density $g(y_{1}|y_{2},\mathbf{x};\boldsymbol{\theta})$,

$$\mathrm{E}[\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})'|y_{i2},\mathbf{x}_{i}] = -\mathrm{E}[\mathbf{H}_{i1}(\boldsymbol{\theta}_{o})|y_{i2},\mathbf{x}_{i}], \tag{13.70}$$

where $\mathbf{H}_{i1}(\boldsymbol{\theta}) = \nabla_{\theta}\mathbf{s}_{i1}(\boldsymbol{\theta})$. Since r_{i2} is a function of $(y_{i2},\mathbf{x}_{i})$, we can put r_{i2} inside both expectations in (13.70). Then, by iterated expectatins,

$$\mathrm{E}[r_{i2}\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})\mathbf{s}_{i1}(\boldsymbol{\theta}_{o})'] = -\mathrm{E}[r_{i2}\mathbf{H}_{i1}(\boldsymbol{\theta}_{o})].$$

Combining all the pieces, we have shown that

$$\begin{aligned}\mathrm{E}[\mathbf{s}_{i}(\boldsymbol{\theta}_{o})\mathbf{s}_{i}(\boldsymbol{\theta}_{o})'] &= -\mathrm{E}[r_{i2}\mathbf{H}_{i1}(\boldsymbol{\theta}_{o})] - \mathrm{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_{o})] \\ &= -\{\mathrm{E}[r_{i2}\nabla_{\theta}\mathbf{s}_{i1}(\boldsymbol{\theta}) + \nabla_{\theta}\mathbf{s}_{i2}(\boldsymbol{\theta})] \\ &= -\mathrm{E}[\nabla_{\theta}^{2}\ell_{i}(\boldsymbol{\theta})] \equiv -\mathrm{E}[\mathbf{H}_{i}(\boldsymbol{\theta})].\end{aligned}$$

So we have verified that an unconditional IM equality holds, which means we can estimate the asymptotic variance of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_{o})$ by $\{-\mathrm{E}[\mathbf{H}_{i}(\boldsymbol{\theta})]\}^{-1}$.

d. From part c, one consistent estimator of $\sqrt{N}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_o)$ is

$$N^{-1}\sum_{i=1}^{N}(r_{i2}\hat{\mathbf{H}}_{i1} + \hat{\mathbf{H}}_{i2}),$$

where the notation should be obvious. But, as we discussed in Chapters 12 and 13, this estimator need not be positive definite. Instead, we can break the problem into needed consistent estimators of $-\mathrm{E}[r_{i2}\mathbf{H}_{i1}(\boldsymbol{\theta}_o)]$ and $-\mathrm{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o)]$, for which we can use iterated expectations. Since, by definition, $\mathbf{A}_{i2}(\boldsymbol{\theta}_o) \equiv -\mathrm{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o)|\mathbf{x}_i]$, $N^{-1}\sum_{i=1}^{N}\hat{\mathbf{A}}_{i2}$ is consistent for $-\mathrm{E}[\mathbf{H}_{i2}(\boldsymbol{\theta}_o)]$ by the usual iterated expectations argument. Similarly, since $\mathbf{A}_{i1}(\boldsymbol{\theta}_o) \equiv -\mathrm{E}[\mathbf{H}_{i1}(\boldsymbol{\theta}_o)|y_{i2},\mathbf{x}_i]$, and r_{i2} is a function of $(y_{i2},\mathbf{x}_i)$, it follows that $\mathrm{E}[r_{i2}\mathbf{A}_{i1}(\boldsymbol{\theta}_o)] = -\mathrm{E}[r_{i2}\mathbf{H}_{i1}(\boldsymbol{\theta}_o)]$. This implies that, under general regularity conditions, $N^{-1}\sum_{i=1}^{N}r_{i2}\hat{\mathbf{A}}_{i1}$ consistently estimates $-\mathrm{E}[r_{i2}\mathbf{H}_{i1}(\boldsymbol{\theta}_o)]$. This completes what we needed to show. Interestingly, even though we do not have a true conditional maximum likelihood problem, we can still used the conditional expectations of the hessians -- but conditioned on different sets of variables, $(y_{i2},\mathbf{x}_i)$ in one case, and $\mathbf{x}_i$ in the other -- to consistently estimate the asymptotic variance of the partial MLE.

e. Bonus Question: Show that if we were able to use the entire random sample, the result conditional MLE would be more efficient than the partial MLE based on the selected sample.

Answer: We use a basic fact about positive definite matrices: if $\mathbf{A}$ and $\mathbf{B}$ are $P \times P$ positive definite matrices, then $\mathbf{A} - \mathbf{B}$ is p.s.d. if and only if $\mathbf{B}^{-1} - \mathbf{A}^{-1}$ is positive definite. Now, as we showed in part d, the asymptotic variance of the partial MLE is $\{\mathrm{E}[r_{i2}\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)]\}^{-1}$. If we could use the entire random sample for both terms, the asymptotic variance would be $\{\mathrm{E}[\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)]\}^{-1}$. But $\{\mathrm{E}[r_{i2}\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)]\}^{-1} - \{\mathrm{E}[\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)]\}^{-1}$ is p.s.d. because $\mathrm{E}[\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)] - \mathrm{E}[r_{i2}\mathbf{A}_{i1}(\boldsymbol{\theta}_o) + \mathbf{A}_{i2}(\boldsymbol{\theta}_o)]$

$= \mathrm{E}[(1 - r_{i2})\mathbf{A}_{i1}(\boldsymbol{\theta}_o)]$ is p.s.d. (since $\mathbf{A}_{i1}(\boldsymbol{\theta}_o)$ is p.s.d. and $1 - r_{i2} \geq 0$.

13.9. To be added.

13.11. To be added.

CHAPTER 14

14.1. a. The simplest way to estimate (14.35) is by 2SLS, using instruments $(\mathbf{x}_1,\mathbf{x}_2)$. Nonlinear functions of these can be added to the instrument list -- these would generally improve efficiency if $\gamma_2 \neq 1$. If $\mathrm{E}(u_2^2|\mathbf{x}) = \sigma_2^2$, 2SLS using the given list of instruments is the efficient, single equation GMM estimator. Otherwise, the optimal weighting matrix that allows heteroskedasticity of unknown form should be used. Finally, one could try to use the optimal instruments derived in section 14.5.3. Even under homoskedasticity, these are difficult, if not impossible, to find analytically if $\gamma_2 \neq 1$.

b. No. If $\gamma_1 = 0$, the parameter γ_2 does not appear in the model. Of course, if we knew $\gamma_1 = 0$, we would consistently estimate $\boldsymbol{\delta}_1$ by OLS.

c. We can see this by obtaining $\mathrm{E}(y_1|\mathbf{x})$:

$$\begin{aligned}\mathrm{E}(y_1|\mathbf{x}) &= \mathbf{x}_1\boldsymbol{\delta}_1 + \gamma_1\mathrm{E}(y_2^{\gamma_2}|\mathbf{x}) + \mathrm{E}(u_1|\mathbf{x}) \\ &= \mathbf{x}_1\boldsymbol{\delta}_1 + \gamma_1\mathrm{E}(y_2^{\gamma_2}|\mathbf{x}).\end{aligned}$$

Now, when $\gamma_2 \neq 1$, $\mathrm{E}(y_2^{\gamma_2}|\mathbf{x}) \neq [\mathrm{E}(y_2|\mathbf{x})]^{\gamma_2}$, so we cannot write

$$\mathrm{E}(y_1|\mathbf{x}) = \mathbf{x}_1\boldsymbol{\delta}_1 + \gamma_1(\mathbf{x}\boldsymbol{\delta}_2)^{\gamma_2};$$

in fact, we cannot find $\mathrm{E}(y_1|\mathbf{x})$ without more assumptions. While the regression y_2 on $\mathbf{x}_2$ consistently estimates $\boldsymbol{\delta}_2$, the two-step NLS estimator of

y_{i1} on $\mathbf{x}_{i1}$, $(\mathbf{x}_i\hat{\boldsymbol{\delta}}_2)^{\gamma_2}$ will not be consistent for $\boldsymbol{\delta}_1$ and γ_2. (This is an example of a "forbidden regression.") When $\gamma_2 = 1$, the plug-in method works: it is just the usual 2SLS estimator.

14.3. Let $\mathbf{Z}_i^*$ be the $G \times G$ matrix of optimal instruments in (14.63), where we suppress its dependence on $\mathbf{x}_i$. Let $\mathbf{Z}_i$ be a $G \times L$ matrix that is a function of $\mathbf{x}_i$ and let Ξ_o be the probability limit of the weighting matrix. Then the asymptotic variance of the GMM estimator has the form (14.10) with $\mathbf{G}_o = \mathrm{E}[\mathbf{Z}_i'\mathbf{R}_o(\mathbf{x}_i)]$. So, in (14.54), take $\mathbf{A} \equiv \mathbf{G}_o'\Xi_o\mathbf{G}_o$ and $\mathbf{s}(\mathbf{w}_i) \equiv \mathbf{G}_o'\Xi_o\mathbf{Z}_i'\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta}_o)$. The optimal score function is $\mathbf{s}^*(\mathbf{w}_i) \equiv \mathbf{R}_o(\mathbf{x}_i)'\boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1}\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta}_o)$. Now we can verify (14.57) with $\rho = 1$:

$$\begin{aligned}\mathrm{E}[\mathbf{s}(\mathbf{w}_i)\mathbf{s}^*(\mathbf{w}_i)'] &= \mathbf{G}_o'\Xi_o\mathrm{E}[\mathbf{Z}_i'\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta}_o)\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta}_o)'\boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1}\mathbf{R}_o(\mathbf{x}_i)] \\ &= \mathbf{G}_o'\Xi_o\mathrm{E}[\mathbf{Z}_i'\mathrm{E}\{\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta}_o)\mathbf{r}(\mathbf{w}_i,\boldsymbol{\theta}_o)'|\mathbf{x}_i\}\boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1}\mathbf{R}_o(\mathbf{x}_i)] \\ &= \mathbf{G}_o'\Xi_o\mathrm{E}[\mathbf{Z}_i'\boldsymbol{\Omega}_o(\mathbf{x}_i)\boldsymbol{\Omega}_o(\mathbf{x}_i)^{-1}\mathbf{R}_o(\mathbf{x}_i)] = \mathbf{G}_o'\Xi_o\mathbf{G}_o = \mathbf{A}.\end{aligned}$$

14.5. We can write the unrestricted linear projection as

$$y_{it} = \pi_{t0} + \mathbf{x}_i\boldsymbol{\pi}_t + v_{it},\ t = 1,2,3,$$

where $\boldsymbol{\pi}_t$ is $1 + 3K \times 1$, and then $\boldsymbol{\pi}$ is the $3 + 9K \times 1$ vector obtained by stacking the $\boldsymbol{\pi}_t$. Let $\boldsymbol{\theta} = (\psi,\boldsymbol{\lambda}_1',\boldsymbol{\lambda}_2',\boldsymbol{\lambda}_3',\boldsymbol{\beta}')'$. With the restrictions imposed on the $\boldsymbol{\pi}_t$ we have

$$\pi_{t0} = \psi,\ t = 1,2,3,\ \boldsymbol{\pi}_1 = [(\boldsymbol{\lambda}_1 + \boldsymbol{\beta})',\boldsymbol{\lambda}_2',\boldsymbol{\lambda}_3']',$$
$$\boldsymbol{\pi}_2 = [\boldsymbol{\lambda}_1',(\boldsymbol{\lambda}_2 + \boldsymbol{\beta})',\boldsymbol{\lambda}_3']',\ \boldsymbol{\pi}_3 = [\boldsymbol{\lambda}_1',\boldsymbol{\lambda}_2',(\boldsymbol{\lambda}_3 + \boldsymbol{\beta})']'.$$

Therefore, we can write $\boldsymbol{\pi} = \mathbf{H}\boldsymbol{\theta}$ for the $(3 + 9K) \times (1 + 4K)$ matrix $\mathbf{H}$ defined by

$$\mathbf{H} = \begin{pmatrix} 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \mathbf{I}_K \\ 0 & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} \\ 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{I}_K \\ 0 & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{0} \\ 1 & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{I}_K & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{I}_K & \mathbf{0} & \mathbf{0} \\ 0 & \mathbf{0} & \mathbf{0} & \mathbf{I}_K & \mathbf{I}_K \end{pmatrix}.$$

14.7. With $\mathbf{h}(\boldsymbol{\theta}) = \mathbf{H}\boldsymbol{\theta}$, the minimization problem becomes

$$\min_{\boldsymbol{\theta}\in\mathbb{R}^P} (\hat{\boldsymbol{\pi}} - \mathbf{H}\boldsymbol{\theta})'\hat{\Xi}^{-1}(\hat{\boldsymbol{\pi}} - \mathbf{H}\boldsymbol{\theta}),$$

where it is assumed that no restrictions are placed on $\boldsymbol{\theta}$. The first order condition is easily seen to be

$$-2\mathbf{H}'\hat{\Xi}^{-1}(\hat{\boldsymbol{\pi}} - \mathbf{H}\hat{\boldsymbol{\theta}}) = \mathbf{0} \quad \text{or} \quad (\mathbf{H}'\hat{\Xi}^{-1}\mathbf{H})\hat{\boldsymbol{\theta}} = \mathbf{H}'\hat{\Xi}^{-1}\hat{\boldsymbol{\pi}}.$$

Therefore, assuming $\mathbf{H}'\hat{\Xi}^{-1}\mathbf{H}$ is nonsingular -- which occurs w.p.a.1. when $\mathbf{H}'\Xi_o^{-1}\mathbf{H}$ -- is nonsingular -- we have $\hat{\boldsymbol{\theta}} = (\mathbf{H}'\hat{\Xi}^{-1}\mathbf{H})^{-1}\mathbf{H}'\hat{\Xi}^{-1}\hat{\boldsymbol{\pi}}$.

14.9. We have to verify equations (14.55) and (14.56) for the random effects and fixed effects estimators. The choices of $\mathbf{s}_{i1}$, $\mathbf{s}_{i2}$ (with added i subscripts for clarity), $\mathbf{A}_1$, and $\mathbf{A}_2$ are given in the hint. Now, from Chapter 10, we know that $\mathrm{E}(\mathbf{r}_i\mathbf{r}_i'|\mathbf{x}_i) = \sigma_u^2\mathbf{I}_T$ under RE.1, RE.2, and RE.3, where $\mathbf{r}_i = \mathbf{v}_i - \lambda\mathbf{j}_T\bar{v}_i$. Therefore, $\mathrm{E}(\mathbf{s}_{i1}\mathbf{s}_{i1}') = \mathrm{E}(\check{\mathbf{X}}_i'\mathbf{r}_i\mathbf{r}_i'\check{\mathbf{X}}_i) = \sigma_u^2\mathrm{E}(\check{\mathbf{X}}_i'\check{\mathbf{X}}_i) \equiv \sigma_u^2\mathbf{A}_1$ by the usual iterated expectations argument. This means that, in (14.55), $\rho \equiv \sigma_u^2$. Now, we just need to verify (14.56) for this choice of ρ. But $\mathbf{s}_{i2}\mathbf{s}_{i1}' = \ddot{\mathbf{X}}_i'\mathbf{u}_i\mathbf{r}_i'\check{\mathbf{X}}_i$. Now, as described in the hint, $\ddot{\mathbf{X}}_i'\mathbf{r}_i = \ddot{\mathbf{X}}_i'(\mathbf{v}_i - \lambda\mathbf{j}_T\bar{v}_i) = \ddot{\mathbf{X}}_i'\mathbf{v}_i = \ddot{\mathbf{X}}_i'(c_i\mathbf{j}_T + \mathbf{u}_i) =$

$\ddot{\mathbf{X}}_i'\mathbf{u}_i$. So $\mathbf{s}_{i2}\mathbf{s}_{i1}' = \ddot{\mathbf{X}}_i'\mathbf{r}_i\mathbf{r}_i'\check{\mathbf{X}}_i$ and therefore $E(\mathbf{s}_{i2}\mathbf{s}_{i1}'|\mathbf{x}_i) = \ddot{\mathbf{X}}_i'E(\mathbf{r}_i\mathbf{r}_i'|\mathbf{x}_i)\check{\mathbf{X}}_i = \sigma_u^2\ddot{\mathbf{X}}_i'\check{\mathbf{X}}_i$. It follows that $E(\mathbf{s}_{i2}\mathbf{s}_{i1}') = \sigma_u^2 E(\ddot{\mathbf{X}}_i'\check{\mathbf{X}}_i)$. To finish off the proof, note that $\ddot{\mathbf{X}}_i'\check{\mathbf{X}}_i = \ddot{\mathbf{X}}_i'(\mathbf{X}_i - \lambda\mathbf{j}_T\bar{\mathbf{x}}_i) = \ddot{\mathbf{X}}_i'\mathbf{X}_i = \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i$. This verifies (14.56) with $\rho = \sigma_u^2$.